National Climatic Data Center

DATA DOCUMENTATION

FOR

DATA SET 9101 (DSI-9101)

Global Daily Climatology Network, V1.0

December 16, 2002

National Climatic Data Center
151 Patton Ave.
Asheville, NC 28801-5001 USA

:

Table of Contents

Topic	Page Number
1. Abstract	3
2. Element Names and Definitions:	3
3. Start Date	6
4. Stop Date	6
5. Coverage	6
6. How to order data	6
7. Archiving Data Center.	6
8. Technical Contact	6
9. Known Uncorrected Problems	6
10. Quality Statement	6
11. Essential Companion Data Sets	10
12. References	10
APPENDIX A	11
APPENDIX B	13
APPENDIX C	14
APPENDIX D	16
APPENDIX E	17
APPENDIX F	18

:
:

1. **Abstract**: The Global Daily Climatology Network (GDCN) represents a compilation of global daily timescale data into a single and consistent format. The data set will serve the needs of researchers, weather-sensitive businesses, agriculture, and policy makers whose are dependent upon complete and accurate analysis of daily temperature and precipitation. Data within the GDCN have been extensively checked through a series of quality control procedures to ensure erroneous values have been removed and/or identified. The GDCN currently has over 800 million days of weather data and that number is expected to increase as the GDCN expands in the future.

2. Element Names and Definitions:

Access Method for Data Files:

All elements available for a single station are stored in an ASCII format data file. These data files can be accessed using a variety of programming languages and/or software (e.g. spreadsheets). The data format (per one line) is as follows:

Variable	Type	Width	Start Column	End Column
STNID	Character	11	1	11
YEAR	Integer	4	12	15
MONTH	Integer	2	16	17
ELEMENT	Character	4	18	21
DATA	Integer	5	22	26
FLAG1	Character	1	27	27
FLAG2	Character	1	28	28

NOTE: DATA, FLAG1, and FLAG2 repeat 31 times for a single record (e.g. one line). Therefore one record of data represents one month of daily data.

Access Method for Inventory Files:

Station information can be found in the ASCII format inventory file. The format of the inventory file is as follows (in Fortran variable notation):

Variable	Type	Width	Start Column	End Column
STNID	Character	11	1	11
LATITUDE	Real	7	12	18
LONGITUDE	Real	8	19	26
ELEVATION	Integer	5	27	31
ELEVATION ESTIMATE FLAG	Character	1	32	32
DATA SOURCE FLAG	Character	1	34	34
STN_NAME	Character	30	36	65
WMOID	Integer	5	66	70
STNID2	Character	8	71	78
STNID3	Character	8	79	86

Element Names and Definitions for Data Files

:

:

STATION ID: The 11 character alphanumeric station identification element is comprised of 2 parts ... a 3-digit country code (see Appendix A) and an 8-digit station code. In some cases the last 5 digits of the 8-digit station code are the 5-digit World Meteorological Organization (WMO) number.

Many countries operate weather stations in foreign lands (e.g. possessions, military bases, research stations, etc.) and therefore it is possible for station IDs to have a 3-digit country code that differs from the country where the station is physically located (e.g. Canadian military base located in Europe).

YEAR: 4-digit year

MONTH: 2-digit month

ELEMENT: 4 character weather element that is one of TMAX (max temperature), TMIN (min temperature), or PRCP (total 24 hour precipitation).

DATA (31): There are 31 daily data values (per record) expressed as a 5-digit integer. The missing value is -9999 and non-missing values are in tenths of a degree Celsius for temperature and tenths of a millimeter for precipitation. Therefore, to obtain the whole measurement unit divide the non-missing integer datum by 10.0 to express value as either whole degree Celsius (to the nearest tenth) or whole millimeters (to the nearest tenth).

Example: (1) TMAX value in 5-digit integer format A00302" would be expressed as 30.2 degrees Celsius. (2) PRCP value in 5-digit integer format A01200" would be expressed as 120.0 millimeters.

DATA MEASUREMENT FLAG(31): There are 31 data measurement flags that are expressed as 1 character. These flags occur immediately following each 5-digit integer datum. Data measurement flags are as follows (Note: flags can apply to Tx = Maximum Temperature, Tn = Minimum Temperature, Pr = Total Precipitation):

- ‘’ - Blank = no data measurement information (Tx,Tn,Pr)
- A - accumulated precipitation value (Pr)
- C - begin or continuing precipitation accumulation (Pr)
- E - estimated value provided by source (Pr)
- H - accumulated precipitation indicated by source, however no accumulated precipitation value was provided (Pr)
- J - source provided accumulated precipitation value, however no information on accumulation period was provided (Pr)
- T - trace accumulation reported by source (Pr)

QUALITY CONTROL FLAG(31): There are 31 quality control flags that are expressed as 1 character. These flags occur immediately following each data measurement flag. The quality control flags are as follows (Note: flags can apply to Tx = Maximum Temperature, Tn = Minimum Temperature, Pr = Total Precipitation):

- ‘’ - Blank = no quality control information (Tx,Tn,Pr)
- F - Failed source quality control (Tx,Tn,Pr)
- S - Suspect value identified by source(Tx,Tn,Pr)
- X - Exceeds known world daily extreme for that particular element (Krause and Flood, 1997). (Tx,Tn,Pr)

:
:

O - Temperature value is greater than or equal to 6.0 bi-weight standard deviations from the period biweight mean. (Tx,Tn)
5 - Temperature value is greater than or equal to 5.0 and less than 6.0 bi-weight standard deviations from the period biweight mean.
4 - Temperature value is greater than or equal to 4.0 and less than 5.0 bi-weight standard deviations from the period bi-weight mean. (Tx,Tn)
3 - Temperature value is greater than or equal to 3.0 and less than 4.0 bi-weight standard deviations from the period bi-weight mean. (Tx,Tn)
K - Streak value (value occurs 10 or more days in a row)

Note: 0.0 mm precipitation values are excluded. (Tx, Tn, Pr))

E - Extreme precipitation value (>= 90th percentile of all observations within that month for the period of record). Example: Jan. 13, 1987, data measurement flag = 'E'. Therefore Jan. 13th 1987 is >= 90th percentile value for all January observations in the entire station record. (Pr)
I - Internal consistency error. The minimum temperature value is greater than same day maximum temperature value. (Tx,Tn)

It is possible that one or more quality control flags may apply to a single datum. Therefore, the above listed flags are listed in order of precedence. Thus, if a value is suspect (e.g. 'S') and also an outlier (e.g. 'O') then the 'S' flag will be set for that datum.

Element Names and Definitions for Inventory File

STATION ID: same as for data file (11 character alphanumeric)

LATITUDE: Expressed in decimal degrees to the nearest hundredth of a degree. (7 digits)

LONGITUDE: Expressed in decimal degrees to the nearest hundredth of a degree. (8 digits)

ELEVATION: Expressed in whole meters (5 digits)

ELEVATION ESTIMATE FLAG: If a station had incomplete elevation information, attempts were made to estimate the elevation based on longitude and latitude. The Global Land One-km Base Elevation (GLOBE) data set was used to estimate the elevation (Hastings et. al., 1997). If left blank this field indicates the elevation was provided in the supplied metadata. If the elevation was estimated using GLOBE then this flag is set to 'E'. (1 character)

DATA SOURCE FLAG: A one-character identifier indicating the source data network from which the data were supplied. In GDCN 1.0 this flag is either blank or set to 'G' to indicate Global Climate Observing System Network stations (GCOS) (Peterson et. al., 1997). (1 character)

STATION NAME: Name of place or location of station. (30 characters)

WMO ID: If supplied in the original metadata this 5 digit integer represents the World Meteorological Organization=s 5 digit station identifier. (5 digits)

:

STATION ID #2: 2nd supplied identifier from original metadata. This often is a cross reference with another station identifier. (8 characters)

STATION ID #3: 3rd supplied identifier from original metadata. This often is a cross reference with another station identifier. (8 characters)

3. Start Date: 18400301

4. Stop Date: 20011130

5. Coverage: Global

6. How to Order Data:

Ask NCDC's Climate Services about the cost of obtaining this data set.
Phone: 828-271-4800
FAX: 828-271-4876
E-mail: **NCDC.Orders@noaa.gov**

7. Archiving Data Center:

National Climatic Data Center
Federal Building
151 Patton Avenue
Asheville, NC 28801-5001
Phone: (828) 271-4800.

8. Technical Contact:

National Climatic Data Center
Federal Building
151 Patton Avenue
Asheville, NC 28801-5001
Phone: (828) 271-4800.

9. Known Uncorrected Problems:

An undetermined number of stations within Mexico have recorded some of their temperature observations in Fahrenheit. No pattern has yet been identified, so users should pay particular attention to the temperature quality control flags. These flags may or may not be useful in identifying conversion issues.

Stations with non-standard practices for recording accumulated precipitation have been identified in Appendix D.

10. Quality Statement: All of the GDCN data (metadata and data) have been processed through an extensive set of quality control procedures. The quality control consisted of two parts:

(1) Simple datum checks (e.g. correct format, impossible values, out of range values, etc.).

(2) Statistical analysis of sets of observations to locate and identify potential outliers and/or erroneous data.

:
:

The user is referred to Appendix C for a complete list of all quality control procedures used in step 1 (above). More extensive explanation for step 2 is provided below for both temperature (maximum and minimum) and precipitation.

Temperature QC:

The quality control of temperature identifies or removes the numerous problems that can exist with raw temperature data. The most notable problems with temperature data are the presence of outliers. Outliers often provoke controversy because when removed they can affect the final analysis of extreme weather events. However if erroneous outliers are included in the data set they can often adversely affect statistical computations. The general philosophy behind GDCN 1.0 is to provide the user with a minimal set of flags to identify the position of the datum within the overall distribution of observations. The user can then make the decision to include the datum based upon their particular application.

The method employed in the identification of outliers is based on robust and resistant measures of location and scale after Lanzante, 1995. The identification of outlier values within GDCN begins with quality control procedure #5 (Appendix C). This procedure appropriately flags all data values greater than or less than the known world extremes as published in the U.S. Army Corps of Engineers (Krause, 1997). Those values that are flagged are then excluded from future quality control calculations. Next, for both maximum and minimum temperature for each day within the station record, a series (X_i) of daily observations is created. This series is comprised of all non-missing data and for a specified day includes the observation from the day before, day after, and the day after, day of, and day before for all other years in the station record. For example, the calculation of the series (X_i) for June 6th, 1988 for a particular station with 3 years of data (1987 through 1989) would consist of the following daily observations:

June 5,6,7 of 1987
June 5,7 of 1988
June 5,6,7 of 1989

This series is used to calculate the biweight mean and biweight estimate of standard deviation. Both of these estimators are more heavily weighted towards the center of their distributions than the tails. Thus, they are more resistant to outlier values. These estimators provide a more robust calculation of mean and standard deviation than traditional statistical methods which apply equal weight throughout the distribution. All stations for which this type of outlier identification was performed required at least 10 years of data so there was at least a minimum of 29 potential observations.

$$u_i = (X_i - M) / (c \times MAD) \qquad (1)$$

A weight (u_i) is calculated for all n observations within (X_i):

where M is he median of X_i and MAD is the median of the absolute deviations of the values from the median (Note: for any $|u_i| \geq 1.0$, $u_i=1.0$). The constant c represents a 'censor' value that controls the distance from the center of the distribution at which the weight is equivalent to zero. The censor value used

:
:

in Lanzante, 1995 of 7.5 was also used throughout all GDCN 1.0 calculations. Therefore the biweight mean from Lanzante, 1995 is:

$$\overline{X}_{bi} = M + \{[\sum_{i=1}^{n}(X_i - M)(1-u_i^2)^2] / \sum_{i=1}^{n}(1-u_i^2)^2\} \qquad (2)$$

and the biweight estimate of standard deviation is:

$$s_{bi} = [n\sum_{i=1}^{n}(X_i - M)^2(1-u_i^2)^4]^{0.5} / \left|\sum_{i=1}^{n}(1-u_i^2)(1-5u_i^2)\right| \qquad (3)$$

The resulting biweight estimators were used to determine the number of biweight standard deviations a particular value was from the biweight mean of the distribution. Values greater than or equal to 3 biweight standard deviations from the biweight mean were classified using the appropriate quality control flags. A summary of these flags is found on pages 6-7 in this documentation. Figure 2 illustrates the identification of these outlier values for a select GDCN 1.0 station.

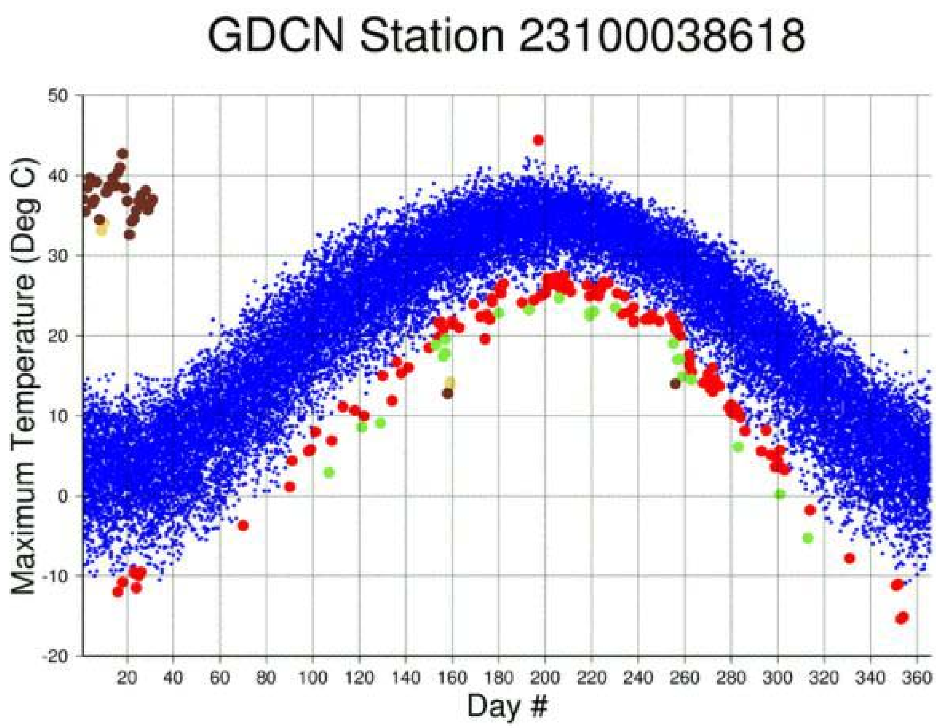

Figure 2. Maximum temperature observations for GDCN station 23100038618. Symbol colors represent quality control flag value - Blue = no flag, Red = '3', Green = '4', Gold = '5', Brown = 'O'.

:

Spatial quality control was conducted for maximum and minimum temperature by comparing monthly averaged values to an independently created gridded data set (Legates and Wilmott, 1990). Each monthly average was compared to the gridded monthly value and a difference series (for the entire period of record) was computed. The median value of this difference series was then computed for each station series and these values were used to determine if a station was retained or excluded from the GDCN 1.0. Based upon this analysis no maximum or minimum temperature stations were excluded from GDCN 1.0.

Precipitation QC:

The quality control of precipitation is often problematic due to the spatial variability of precipitation and inadequate metadata (lack of information concerning observing practices and times). The precipitation quality control consisted of the basic checks listed in Appendix C. In addition careful initial format procedures were implemented to ensure that the received raw precipitation metadata information were preserved within the GDCN data measurement flags. However, this often lead to some discrepancies within the data. The data measurement flags 'H' and 'J' both illustrate two problems unique to accumulated precipitation (see pg. 6). Therefore in some cases the preservation of the incomplete or erroneous supplied metadata allows these problems to persist. Thus, users should pay particular attention to the data measurement flag when processing precipitation data.

Unique observing practices can lead to potentially erroneous analysis conclusions. A careful analysis (by day of week) was performed on all precipitation data to examine potential inconsistencies within the data record. In particular, one non-standard precipitation observing practice was identified. If the accumulated precipitation value was 0.0 mm then the observer would correctly record the accumulation but in the previous days included in the accumulation the observer would also record 0.0 mm values. Although not physically incorrect, these observers did not observe the same practice when the accumulation was greater than 0.0 mm. In those instances the previous days in the accumulated period would be set to the missing value. Furthermore this occurrence was almost exclusive to Monday accumulations from the weekend period. Thus, a careful analysis of the record would erroneously show that Saturday and Sunday always had 0.0 mm of precipitation or a missing value of precipitation, if indeed the general practice was to not observe on the weekend throughout the station record. Fortunately the incidence of this occurrence is small within the GDCN (< 100 stations) and thus Appendix D contains a list of stations that were identified with the above mentioned problem.

Spatial quality control was carried out in a similar manner described in the 'Temperature QC'. No precipitation stations were excluded from the GDCN 1.0 based upon the spatial precipitation analysis.

Duplicate Removal:

With over 800 million days of daily data, direct comparison of all data was not feasible due to computational and time limitations. Therefore, duplicate identification was conducted by comparison of metadata (geographic coordinates). Stations with exact or very similar geographic positions were identified and then their full records were compared and the percent of similarity was computed. Stations with exact or very high similarities were then identified and only the station with the longest total period of record was retained for inclusion in GDCN 1.0.

:

11. Essential Companion Datasets: None.

12. References:

Hastings, David A., and P. K. Dunbar, 1997. The development of global digital elevation data. Proceedings, 18th Conference on Remote Sensing, Kuala Lumpur. pp. J=S-3-1 to J=S-3-6.

Krause, Paul F. and Kathleen Flood. (1997). Weather and Climate Extremes (TEN-0099). Fort Belvoir, VA: U.S. Army Engineer Topographic Engineering Center).

Lanzante, John R. 1996. Resistant, robust, and nonparametric techniques for the analysis of climate data. Theory and examples, including applications to historical radiosonde station data. Int. J Climatol., 16, 1197-1226.

Legates, David R., and Cort J. Willmott. 1990. "Mean seasonal and spatial variability in gaugencorrected global precipitation." International Journal of Climatology, vol. 10. pp. 111n127.

Legates, David R. and Cort J. Willmott. 1990. "Mean seasonal and spatial variability in global surface air temperature." Theoretical and Applied Climatology, vol. 41, pp. 11n21.

Peterson, Thomas, Harald Daan, Philip Jones, 1997: Initial Selection of a GCOS Surface Network. Bulletin of the American Meteorological Society: Vol. 78, No. 10, pp. 2145B2152.

:

Appendix A

3-digit country codes (used as the first 3 digits of the 11 character station ID) Note: A digital file of the following list is available at the GDCN web site - www.ncdc.noaa.gov/gdcn.html under 'Supplementary Files':

101 - Algeria
115 - Egypt
141 - South Africa
148 - Sudan
205 - China
207 - India
210 - Japan
213 - Kyrgyzstan
215 - Mongolia
221 - Korea
222 - Russia
223 - Saudi Arabia
227 - Tajikistan
228 - Thailand
229 - Turkmenistan
231 - Uzbekistan
301 - Argentina
303 - Brazil
306 - Ecuador
308 - Paraguay
313 - Uruguay
314 - Venezuela
403 - Canada
405 - Costa Rica
408 - El Salvador
410 - Guatemala
412 - Honduras
414 - Mexico
415 - Nicaragua
416 - Panama
425 - United States of America
501 - Australia
505 - Malaysia
514 - French Polynesia
531 - New Caledonia
602 - Armenia
603 - Austria
604 - Azerbaijan
605 - Belarus
606 - Belgium
607 - Bosnia
612 - Denmark
613 - Estonia
614 - Finland
615 - France
616 - Georgia
617 - Germany
618 - Greece
620 - Iceland

:
:

621 - Ireland
622 - Israel
623 - Italy
625 - Kazakhstan
626 - Latvia
628 - Lithuania
629 - Luxembourg
631 - Moldova
633 - Netherlands
634 - Norway
635 - Poland
636 - Portugal
641 - Slovakia
642 - Slovenia
643 - Spain
645 - Sweden
646 - Switzerland
647 - Syria
650 - Ukraine
651 - United Kingdom

Appendix B

The National Climatic Data Center staff listed below were responsible for obtaining GDCN data. All data contained within the GDCN are believed to be available for re-distribution through the GDCN in compliance with WMO resolution 40. If you have further questions about the National Climatic Data Center's right to re-distribute data through the GDCN, please email your questions and/or concerns to:

NCDC.GDCN@noaa.gov

Dr. Pavel Ya. Groisman
Visiting UCAR Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Dr. Thomas C. Peterson
Physical Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Dr. David R. Easterling
Principal Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Mr. Russell S. Vose
Chief, Climate Analysis Branch
National Climatic Data Center
Asheville, NC 28801-5001

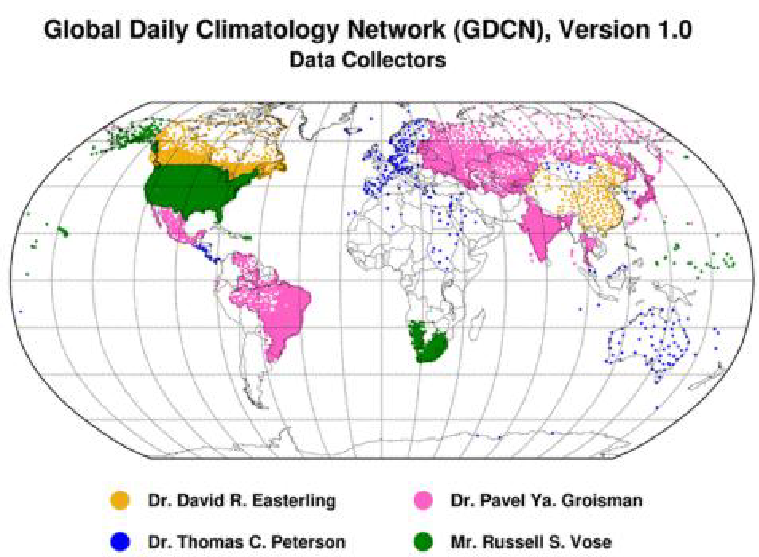

Figure 3. Original GDCN data source collectors

Appendix C

Data and station information format quality control:

1. Check if station identification in station file matches the station identification used in the station file name.

2. Check all non-existent days for missing data. (Example: February, 30th should always equal the missing value)

3. Check for proper sort of station data and inventory file.

4. Trace precipitation (Flag 'T') must equal 0.0.

5. Check that all record elements are one of PRCP (total precipitation), TMAX (maximum temperature), or TMIN (minimum temperature)

6. Check if the 4-digit year element within the station data file is: 1800 <= YEAR <= 2001.

7. Check that month field is 1 <= MONTH <= 12.

8. Check both the data measurement and quality control flags to ensure they only contain documented flags.

9. Check each datum to ensure they are integer values.

10. Check for duplicate stations within the entire data set.

11. Check for station -90.0 <= Latitude <= 90.0
    -180.0 <= Longitude <= 180.0

12. Check that the number of stations within the inventory file matches the number of station data files.

13. Check for precipitation < 0.0 mm.

14. Check for maximum temperature < minimum temperature for the same day.

15. Check for days with the same value repeated 10 or more consecutive times (e.g. quality control flag = 'K'). Note: 0.0 mm precipitation values were excluded.

16. Check for daily element value exceeding known world extreme according to the U.S. Army Corps of Engineers World Extremes (Krause and Flood, 1997).

World Extremes used for this data set:

highest temperature: El Azizia, Libya 57.8 deg C, September 13, 1922

lowest temperature: Vostok, Antarctica -89.4 deg C, July 21, 1983

greatest 24-hr rainfall: Foc-Foc, Réunion Island 1828.8 mm, January 7th to 8th, 1966

17. If source country supplied quality control information (generally failed or suspect quality control) for a datum then the quality control flag was set as either ‘F’ for failed quality control or ‘S’ for suspect datum.

18. Compute monthly sums and/or averages for each station and compare to the global gridded climatology by Legates and Wilmott, 1990.

Appendix D

Stations with potentially inconsistent accumulated precipitation observing practices. See page 14 for further explanation. (Note: A digital file containing a list of the following stations is available at the GDCN web site (www.ncdc.noaa.gov/gdcn.html) under 'Supplementary Files':

14100174520	31300001751	42500361033	42500667843
14100284150	31300001784	42500367978	42500669466
14100315070	40301191440	42500455387	42500669608
14100322090	40302204000	42500456909	42500914000
14101274850	40304013400	42500510145	50100094380
14101516040	40306044903	42500510211	50100094570
14101521900	40306060773	42500510470	50100094693
14101524820	40307022280	42500511075	50100094784
14102090390	40307063400	42500511339	50100094869
14102400730	40307074240	42500511665	50100094910
14102985120	40308100500	42500511887	
14103044460	40308202300	42500512751	
14103688310	40308206450	42500513734	
14103728520	42500055507	42500513982	
14103734850	42500058064	42500514318	
14104002030	42500067432	42500514634	
14104753700	42500105708	42500514742	
14104757610	42500113595	42500515286	
14105116720	42500244328	42500515710	
14105140100	42500300870	42500515864	
14105558780	42500301265	42500516134	
14105945390	42500301580	42500516588	
14105945900	42500304207	42500518760	
14105948060	42500304944	42500518815	
20704171200	42500305171	42500518941	
3030C5-0500	42500305679	42500519955	
3030E3-0530	42500318778	42500660040	
31300001220	42500336786	42500660152	
31300001440	42500338366	42500661345	
31300001558	42500340394	42500662801	
31300001709	42500358780	42500663904	
31300001734	42500360785	42500665020	

:

Appendix E

Global Daily Climatology Network 1.0 Use and Distribution LICENSE

:
:

Appendix F

Acknowledgments:

Below is a list (not necessarily complete) of people and/or institutions who assisted in the construction of this data set by either directly or indirectly providing data and/or advice.

Mr. Red Ezell
Computer Programmer
National Climatic Data Center
Asheville, NC 28801-5001

Dr. David R. Easterling
Principal Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Mr. Russell S. Vose
Chief, Climate Analysis Branch
National Climatic Data Center
Asheville, NC 28801-5001

Mr. Brant Liebmann
Research Scientist
Climate Diagnostics Center
Boulder, CO 80305-3328

Dr. Mark Morrissey
Environmental Verification and
Analysis Center (EVAC)
Norman, OK 73069

ANEEL (Agência Nacional De
Energia Elétrica) - Brazil
www.aneel.gov.br

Mr. Robert Morris
Manager, Information Svcs. Div.
Meteorological Service of Canada
Environment Canada
Toronto, Ontario
M3H 5T4

Dr. Thomas C. Peterson
Physical Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Dr. Pavel Ya. Groisman
Visiting UCAR Scientist
National Climatic Data Center
Asheville, NC 28801-5001

Mr. Roy Jenne
Manager, SCD Data Support Sec.
NCAR
Boulder, CO 80307-3000

Dr. Arthur V. Douglas
Dept. of Atmospheric Science
Creighton University
Omaha, NE 68178

KNMI (Koninklijk Nederlands
Meteorologisch Instituut)
www.knmi.nl/samenw/eca
The Netherlands

National Climate Center
China Meteorological Admin.
Beijing, China

Mr. Johan Koch
South African Weather Service
Pretoria, South Africa

: